A Rebuttal to Zhang's Critique of the Genetic Equidistance Phenomenon and Maximum Genetic Diversity Hypothesis

Shi Huang

Center for Medical Genetics, School of Life Sciences, Central South University, 110 Xiangya Road, Changsha, Hunan 410078, P.R. China

Key words: Genetic equidistance phenomenon (GEP), maximum genetic diversity (MGD) hypothesis, molecular clock, neutral theory

Abstract

In his recent paper, Zhang attempts to refute the genetic equidistance phenomenon (GEP) while defending the molecular clock and neutral theory. However, he overlooks a fundamental contradiction: the neutral theory is based on the molecular clock, which in turn relies on the GEP. While Zhang demonstrates a superficial understanding of common methods in molecular phylogeny, he fails to grasp several critical facts and concepts relevant to the topic. The methods and results presented are largely irrelevant to the GEP, and his criticisms of the maximum genetic diversity (MGD) hypothesis largely amount to a strawman argument. Notably, Zhang neglects the multiple lines of evidence supporting the existence of an upper limit on genetic distance or diversity. He also overlooks the inconvenient fact that the implicit assumption of the molecular clock and neutral theory—that genetic distance has no upper limit—has yet to be validated by any direct tests. As a result, Zhang's paper cannot be considered a meaningful contribution.

The Maximum Genetic Diversity (MGD) theory was published in 2008 (1, 2), inspired by the long-overlooked Genetic Equidistance Phenomenon (GEP), first discovered by Margoliash in 1963(3). The GEP is widely acknowledged to have contributed to the development of the molecular clock hypothesis and is considered to have been well explained by it (4). The molecular clock, in turn, inspired the neutral theory, which is intended to explain or predict the molecular clock, and is considered to be the best evidence for the neutral theory (5). The MGD theory questions the validity of both the molecular clock and the neutral theory, arguing that they misinterpret genetic distance that is already at an upper limit, implicitly assuming it continues to increase with time.

The GEP is reinterpreted by the MGD theory as a result of mutation saturation, where genetic distance reaches a maximum or upper limit. According to this theory, simpler species can tolerate more mutations or higher levels of genetic diversity, which allows them to reach higher saturation levels. The genetic distance between a simpler species and a more complex one is determined by the MGD of the simpler species. As a result, a simpler species exhibits equal genetic differences in its gene sequence as measured by the identity matrix or percentage differences when compared to two or more species of higher complexity. This forms the basis of the GEP.

Recently, Zhang published a preprint arguing that both the GEP and the MGD theory are invalid, while defending the molecular clock and neutral theory (6). In this paper, we refute his arguments. In fact, Zhang demonstrates only a superficial understanding of the relevant literature and the topics at hand. Due to his misunderstanding of the GEP, most of his results are irrelevant in addressing whether the GEP is real. He fails to recognize a fundamental contradiction in his own paper: by attempting to disprove the GEP, he inadvertently undermines the very foundation of the molecular clock and the neutral theory.

In the following, we address the factual and conceptual errors in Zhang's paper in the order they appear.

Page 2, line 1, “In 1963, Margoliash compared hemoglobin sequence”. It was cytochrome c that Margoliash had studied (3).

Page2, line 5-6, “a phenomenon that Huang refers to genetic equidistance”. It was Margoliash who referred to his finding as genetic equidistance or (sequence) “equally different”. The genetic equidistance phenomenon was first noted in 1963 by Margoliash, who wrote: "It appears that the number of residue differences between cytochrome c of any two species is mostly conditioned by the time elapsed since the lines of evolution leading to these two species originally diverged. If this is correct, the cytochrome c of all mammals should be equally different from the cytochrome c of all birds.”(3)

Page 2 line 8,“Huang suggested that the molecular clock was inspired by the phenomenon of genetic equidistance in 1963” Huang actually was not the first to suggest this as implied by Zhang. Margoliash was the first when he suggested time to be the only factor in determining the number of residue difference between two species as shown by the above quote (3). A leading expert Kumar in a review on the history of the molecular clock has also given credit to Margoliash and discussed in length the molecular clock interpretation of the GEP: “This informal proposal of a molecular clock (by Zuckerkandl and Pauling) was followed by a formal statement the following year by Margoliash: It appears that the number of residue differences between cytochrome c of any two species is mostly conditioned by the time elapsed since the lines of evolution leading to these two species originally diverged.”(4). In attempting to refute the GEP

while defending the molecular clock, Zhang apparently failed to recognize the inseparable link between the GEP and the molecular clock.

Page 2, line 9 to 14, “However, there is a difference between the two explications, in that genetic equidistance emphasised the contrast between two closely related “complex” species and one distantly related “simple” species, whereas Zuckerkandl's 1962 paper emphasised that amino acid differences are greater the more distantly related the species.” The findings of Margoliash and Zuckerkandl/Pauling were simply two sides of the same coin, with no differences in the underlying mechanism they each proposed to explain their observations of genetic distance. That mechanism was the molecular clock. Given Zhang's limited understanding of the relevant concepts and literature, it is unsurprising that he perceives differences where none actually exist.

Page 2, first 3 lines, 2nd paragraph, “The earliest molecular clock hypothesis suggested that the rate of mutation was constant and that the timing of divergence could thus be inferred from the number of mutations accumulated in a sequence” This is not completely true. The most critical and controversial aspect of the molecular clock hypothesis is the claim that different species share similar mutation rates. It is surprising that someone writing a paper on the molecular clock would overlook such a fundamental point. Unfortunately, this oversight aligns with the careless approach to scholarship that Zhang has exhibited throughout his paper.

Page 2, last paragraph, “Consider the question of whether the discovery that numerous non-coding regions of the genome are functional absolutely contradicts the neutral theory itself? Obviously not, because functional regions and neutral mutations are not incompatible, and they are selectionally neutral as long as they do not affect the fitness of an individual, since changes in molecular structure or even in the structure of organisms may not affect the fitness associated traits of organisms.” While Zhang correctly pointed out that some DNA sequences may be selectively neutral if the traits they affect do not influence fitness, he overlooked a critical factor: most genes or variants are pleiotropic, influencing multiple traits, and most traits are determined by multiple genes or variants(7). It is highly unlikely that none of the numerous traits associated with a gene or mutation are linked to fitness. For example, a mutation that affects flower color might or might not be under selection, depending on whether it also impacts other traits that are subject to selection, even if we assume that the color change itself does not directly affect fitness. Given that we cannot know in advance whether the color-changing mutation also influences other traits, it is premature to conclude that the mutation is selectively neutral.

Page 3, last sentence of the first paragraph, “It should be stressed, however, that finding fewer neutral loci among genomes is beneficial for phylogenetic analysis, as neutral sites are more likely to be subjected to multiple hits, resulting in wrong tree topology” This is flatly false. It is well-established that phylogenetic trees rely on the neutral assumption. The ideal scenario in phylogenetics is that all sites in a genome are neutral. Additionally, the infinite sites assumption, which is a key part of the neutral framework commonly adopted in the field, posits that multiple mutations at the same site are unlikely. Natural selection could cause two different species to acquire the same adaptive mutations at the same sites, which would disrupt the tree topology.

Page 3 second paragraph. Zhang here misunderstood the MGD theory. He was correct that the field has extensively addressed substitution saturation. Like him, critics of the MGD theory have often pointed out that the saturation issue it raises is not new. However, these critics have failed to recognize the distinction between partial saturation and the upper limit concept central to the MGD. The field has primarily focused on partial saturation, where genetic distance between

species still increases with time, albeit at a nonlinear rate compared to the early stages of species divergence. In a linear model, one mutation results in one residue difference in genetic distance. In a nonlinear model, two or more mutations may result in a single residue difference due to multiple mutations occurring at the same site, which is known as saturation.

The methods the field employs to address saturation assume that genetic distance continues to increase with time, despite partial saturation. This allows for modeling how many mutations have occurred in the past to reach the current level of partial saturation. In contrast, the MGD theory posits that genetic distance has already reached an upper limit, beyond which it is no longer associated with time and the number of mutations. This upper limit could have been reached a long time ago, making it impossible to model the number of mutations that have occurred in the past.

Thus, the key difference between the MGD theory and the prevailing view is that the former considers genetic distance to be at its maximum, while the latter does not. As such, the issue of maximum genetic distance has not been addressed by the field and, in principle, cannot be addressed by existing methods. Zhang failed to recognize this fundamental distinction.

Page 3, third paragraph. “Another important refutation of the molecular clock hypothesis claimed by Huang is the fact that mutational rates vary among genomic regions and clades……. To sum up, advocators of the MGD hypothesis failed to provide a persuasive reason to deny the molecular clock hypothesis.” Zhang fails to recognize that mutation rate variation, as observed in reality where mutation rates among different vertebrate species can vary by up to 40 times (8), directly invalidates the key claim of the molecular clock hypothesis—that all species share similar mutation rates. This may stem from his lack of a proper understanding of the molecular clock hypothesis.

However, there is a more fundamental reason why the molecular clock must be considered invalid. Even if all species did share the same mutation rate, the molecular clock would still be invalid because it misinterprets genetic distance at the upper limit level as continuing to increase over time. It appears, then, that Zhang has not fully grasped either the molecular clock hypothesis or the MGD theory. Therefore, as we have explained above, most of the points Zhang makes in the introduction section of his paper are flawed and factually inaccurate, reflecting his poor scholarship.

In the Methods and Results section, Zhang used the JTT matrix in most of his analyses to measure genetic distance (see supplementary tables 2 and 3). He also applied a gamma distribution to model rate variation (see supplementary tables 4 and 5). The only result presented using the identity matrix is found in supplementary table 1.

There are various sequence similarity scoring matrices. The simplest and most straightforward is the "identity matrix," also referred to as the "Unitary Protein Matrix" (UPM). Other matrices include the Genetic Code Matrix, the Structure-Genetic Matrix, the JTT matrix, and others. The JTT matrix (Jones-Taylor-Thornton matrix) is a substitution matrix based on observed amino acid substitution frequencies in protein evolution (9). It gives a probability of one amino acid being replaced by another, considering the evolutionary history and likelihood of specific substitutions. All of these non-identity matrices assume a particular view of evolutionary history that may or may not be accurate and represent a human-made interpretation or transformation of the basic identity matrix.

The original genetic distance results from Margoliash and Zuckerkandl/Pauling were calculated using the identity matrix, or simply percentage non-identity (for example, if there are 10 residue differences in a 100-amino acid alignment between two species, the distance is 10% difference). The identity matrix is based solely on factual observations and does not involve any human-made assumptions or interpretations. The first question that should be asked is: is the observed distance still increasing with time, or has it reached an upper limit? This question was never addressed until we raised it. Many have implicitly assumed that genetic distance has not reached an upper limit, but our work has proven this assumption to be incorrect.

Therefore, anyone seeking to verify whether the GEP is real should measure genetic distance using the identity matrix rather than any other matrix, as it was first established using the identity matrix. Indeed, we have consistently used the identity matrix in all our papers. Zhang failed to grasp this key point, as he performed most of his analyses using irrelevant methods, such as the JTT matrix and the rate variation modeling with a gamma distribution (4 out of 5 results, including the two figures and supplementary tables 2-5). These results therefore are not relevant to the GEP.

The only relevant result is in supplementary table 1, where the identity matrix was used to score genetic distance in cytochrome c among various species. The problem, however, is that Zhang failed to conduct the analysis properly. As we have previously done for numerous species and genes in demonstrating the GEP (10, 11), a work that Zhang has not acknowledged, he should have first selected a group of at least three species, with the complexity of each species intuitively apparent based on the relative differences in the estimated number of cell types for each species. He should then have compared the two more complex species to the simpler species to determine if they are equidistant to the simpler species.

Unfortunately, Zhang did not approach the analysis this way, and it is unclear which species he considered to be the simpler one among those he chose for analysis. For instance, Zhang compared sheep to human, horse, cattle, and pig in cytochrome c distance and found that sheep is closer to all the ungulates than to humans. However, since human is the outgroup to the ungulates, it is expected that the different ungulates would be closer to each other due to their closer physiology and phylogenetic relationship. A closer genetic distance in a non-neutral gene like cytochrome c could indicate both closer physiological and phylogenetic relationships. Furthermore, it is unclear whether sheep is the least complex among the ungulates. Thus, the result in supplementary table 1 is not informative regarding the GEP issue. It was precisely by using cytochrome c that Margoliash demonstrated the GEP (3). Therefore, Zhang's flawed analysis of the same protein can hardly be considered a serious challenge to Margoliash's work (3) or our work (10, 11). It seems that none of Zhang's results provide useful insights into the question he set out to address.

It should be emphasized that if Zhang had successfully disproven the reality of the GEP, he would have single-handedly overturned the entire field of molecular evolution, including the molecular clock and neutral theory. This is because, if the GEP were untrue, the molecular clock would merely explain an artifact, and the neutral theory would then be an explanation for the fake molecular clock. Zhang's failure to understand the cause-and-effect relationship between the GEP and the molecular clock makes it not only self-defeating but also preposterous for him to attempt to disprove the GEP while simultaneously defending the molecular clock and neutral theory.

In the section titled "What is the Upper Limit Theory," Zhang states: "So if we look at genetic diversity at this large scale, then there exists a limit." This statement aligns with the core concept of the MGD theory. Does this mean that Zhang agrees with the MGD theory? If so, then what exactly is the point of his paper? This contradiction is both confusing and self-defeating. It appears that Zhang has not carefully considered his ideas and has failed to present a self-consistent and coherent body of work.

In the section "MGD or LBA," it is unclear exactly what the author is trying to argue. Zhang failed to understand that while LBA (long-branch attraction) involves mutation saturation, where two species have identical residues that are not due to common ancestry but due to independent mutations, MGD concerns repeated mutations at the same site that lead to a different residue. As explained in our previous paper, saturation in the context of MGD is 20 times more likely than LBA, given there are 20 amino acids (12). When genetic distance reaches its maximum, mutations will still occur but will mostly result in changes at the same site or locus, shifting from one amino acid to a different one. For example, if species 1 has amino acid A and species 2 has amino acid B at a specific site, mutations in species 1 affecting the A residue would most likely result in a new amino acid, different from B. Therefore, the genetic distance would stay unchanged despite the mutations. In contrast, LBA would lead to a reduction in distance, changing A to B. LBA is a special case of reaching saturation. The question that both Zhang and the field should ask is: if LBA is common, shouldn’t MGD be even more common?

In the section titled "Is Macroevolution the Evolution of Species Complexity Increasing?", Zhang argues and provides examples of evolution toward lower complexity, which he believes undermines the thesis of the MGD theory that evolution tends to lead to higher complexity. While "complexity" is difficult to define in general, the MGD theory defines it in terms of the number of cell types, a metric that seems to capture the major transitions in evolution, such as the shift from apes to humans.

Although there are some infrequent cases of apparent loss of complexity, such as the loss of eyesight in nearly blind naked mole rats, these changes are relatively minor and do not detract from the overarching trend of progression toward higher complexity. Moreover, it remains unclear whether there may be compensatory increases in complexity in other aspects, leading to overall minimal net changes in complexity. In any case, the overarching trend in evolution is not from bacteria to bacteria or from humans to bacteria, but rather from bacteria to humans. This supports the idea of a general direction toward higher complexity, rather than a trend toward simpler forms or no direction at all.

Theories proposing a direction toward simplicity have yet to emerge. Popular theories like those of Darwin and Kimura argue for no specific direction but fail to account for the GEP. In contrast, the MGD theory proposes a direction toward higher complexity and offers a coherent explanation for all major evolutionary phenomena. As such, it may represent the best working model available at this time.

Finally, in terms of the key difference between the neutral theory and MGD, Zhang overlooked the lack of any tests to confirm the implicit assumption of the neutral theory that genetic distance is always increasing with time. It is also important to highlight Zhang's failure to consider multiple lines of evidence that have confirmed the MGD theory and invalidated the molecular clock and neutral theory.

1. A substantial number of residues in proteins have undergone mutation saturation, far exceeding the predictions made by the molecular clock and the neutral theory (13). While the molecular clock can superficially account for the numeric value of genetic distance, it fails to explain the oversaturation of mutations observed in the GEP. However, this shortcoming was not fully recognized until we pointed it out.
2. The genetic non-equidistance phenomenon, predicted by the MGD theory, contradicts the molecular clock. For example, while owl monkeys and marmosets—both New World monkeys—should be equidistant to the outgroup human if the molecular clock were correct, the more complex owl monkey is actually closer to humans (14), a result fully anticipated by the MGD theory (15, 16).
3. The genetic diversity of patient populations is greater than the normal controls (17, 18).
4. The overall level of individual genetic diversity, measured by heterozygosity, minor allele contents of SNPs, the number of loss-of-function variants, or the number of non-synonymous variants, is inversely related to cognitive function (19-24).
5. The dramatic difference between fast and slowly evolving genes in genetic diversity (15, 25).

In summary, while Zhang's paper demonstrates a superficial understanding of the field of molecular evolution and some of the technical methodologies commonly used, it is evident that he lacks a coherent and accurate grasp of the fundamental facts, history, and concepts. His understanding of the MGD theory is also incomplete. We suspect that Zhang may not be alone in his confusion, as such misunderstandings could be common among practitioners in the field, given the incoherent nature of the prevailing framework. By attempting to invalidate the GEP while defending the molecular clock and neutral theory, Zhang fails to recognize the self-contradictory and self-defeating nature of his argument. He overlooks multiple lines of evidence confirming that genetic distance has reached an upper limit in most cases, as well as the lack of tests to verify the implicit assumption of the molecular clock that genetic distance always increases over time. As a result, while we would welcome a well-presented paper challenging the GEP and the MGD theory, Zhang's paper is simply too flawed to be considered a meaningful contribution.

Acknowledgements:

This work was not funded by any grant agencies.

References:

1. Huang S. Inverse relationship between genetic diversity and epigenetic complexity. Nature Precedings 2009:doi.org/10.1038/npre.2009.1751.2.
2. Huang S. Histone methylation and the initiation of cancer, Cancer Epigenetics. Tollefsbol T, editor. New York: CRC Press; 2008.
3. Margoliash E. Primary structure and evolution of cytochrome c. Proc Natl Acad Sci 1963;50:672-9.
4. Kumar S. Molecular clocks: four decades of evolution. Nat Rev Genet. 2005;6(8):654-62.

5. Kimura M, Ohta T. Protein polymorphism as a phase of molecular evolution. Nature 1971;229:467-79.
6. Zhang Y. The genetic equidistance and maximum genetic diversity hypothesis: Smoke and mirrors? BioRxiv. 2023:https://doi.org/10.1101/2023.02.14.528494.
7. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177-86.
8. Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, et al. Evolution of the germline mutation rate across vertebrates. Nature. 2023;615(7951):285-91.
9. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 1992;8(3):275-82.
10. Luo D, Huang S. The genetic equidistance phenomenon at the proteomic level. Genomics. 2016;108(1):25-30.
11. Yuan D, Huang S. Genetic equidistance at nucleotide level. Genomics. 2017;109:192-5.
12. Wang M, Wang D, Yu J, Huang S. Enrichment in conservative amino acid changes among fixed and standing missense variations in slowly evolving proteins. PeerJ. 2020;8:e9983 https://doi.org/10.7717/peerj.9983.
13. Huang S. The overlap feature of the genetic equidistance result, a fundamental biological phenomenon overlooked for nearly half of a century. Biological Theory. 2010;5:40-52.
14. Mao Y, Harvey WT, Porubsky D, Munson KM, Hoekzema K, Lewis AP, et al. Structurally divergent and recurrently mutated regions of primate genomes. Cell. 2024;187(6):1547-62 e13.
15. Huang S. Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers. Sci China Life Sci. 2012;55:709-25.
16. Bickel D. Testing Hypotheses of Molecular Evolution. In: Phylogenetic Trees and Molecular Evolution.: Springer, Cham.; 2022.
17. Zhu Z, Yuan D, Luo D, Lu X, Huang S. Enrichment of Minor Alleles of Common SNPs and Improved Risk Prediction for Parkinson's Disease. PLoS ONE. 2015;10(7):e0133421.
18. Huang S. The maximum genetic diversity theory of molecular evolution. Communications in Information and Systems. 2023;23:359-92.
19. Wang M, Huang S. The collective effects of genetic variants and complex traits. J Hum Genet. 2023;68:255-62.
20. Chen CY, Tian R, Ge T, Lam M, Sanchez-Andrade G, Singh T, et al. The impact of rare protein coding genetic variation on adult cognitive function. Nat Genet. 2023;55(6):927-38.
21. Ganna A, Genovese G, Howrigan DP, Byrnes A, Kurki M, Zekavat SM, et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat Neurosci. 2016;19(12):1563-5.
22. Ganna A, Satterstrom FK, Zekavat SM, Das I, Kurki MI, Churchhouse C, et al. Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. Am J Hum Genet. 2018;102(6):1204-11.

328 23. Sha Z, Sun KY, Jung B, Barzilay R, Moore TM, Almasy L, et al. The copy
329 number variant architecture of psychopathology and cognitive development in
330 the ABCD((R)) study. medRxiv. 2024.
331 24. Gardner EJ, Neville MDC, Samocha KE, Barclay K, Kolk M, Niemi MEK, et al.
332 Reduced reproductive success is associated with selective constraint on human
333 genes. Nature. 2022;603(7903):858-63.
334 25. Yuan D, Lei X, Gui Y, Wang M, Zhang Y, Zhu Z, et al. Modern human origins:
335 multiregional evolution of autosomes and East Asia origin of Y and mtDNA.
336 bioRxiv. 2017:doi.org/10.1101/101410

337